The Annals of Statistics
2017, Vol. 45, No. 3, 1312–1341
DOI: 10.1214/16-AOS1486

OPERATIONAL TIME AND IN-SAMPLE DENSITY FORECASTING

BY YOUNG K. LEE[1,*], ENNO MAMMEN[2,†], JENS P. NIELSEN[3,‡] AND BYEONG U. PARK[4,§]

Kangwon National University[*]*, Universität Heidelberg and Higher School of Economics*[†]*, Cass Business School, City University London*[‡] *and Seoul National University*[§]

In this paper, we consider a new structural model for in-sample density forecasting. In-sample density forecasting is to estimate a structured density on a region where data are observed and then reuse the estimated structured density on some region where data are not observed. Our structural assumption is that the density is a product of one-dimensional functions with one function sitting on the scale of a transformed space of observations. The transformation involves another unknown one-dimensional function, so that our model is formulated via a known smooth function of three underlying unknown one-dimensional functions. We present an innovative way of estimating the one-dimensional functions and show that all the estimators of the three components achieve the optimal one-dimensional rate of convergence. We illustrate how one can use our approach by analyzing a real dataset, and also verify the tractable finite sample performance of the method via a simulation study.

1. Introduction. In-sample forecasting is a recently introduced class of forecasting methods based on structured nonparametric models. The idea is that observations might fall in some set, say S, in $\mathbb{R}^2$ and that S can be written as the union of two subsets S_1 and S_2, where S_1 is the set of observed observations and S_2 is the set of future observations whose distribution is the target for forecasting. In-sample density forecasting assumes that the density restricted to S_1 or to S_2 can be described by the same one-dimensional nonparametric functions. This assumptions leads to the convenient forecasting strategy of estimating the structured density on the observed data in S_1 and then simply reusing the nonparametrically estimated

Received July 2015; revised June 2016.

[1]Supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2015R1A2A2A01005039).

[2]Supported by Deutsche Forschungsgemeinschaft through the Research Training Group RTG 1953 and by the Government of the Russian Federation within the framework of the implementation of the Global Competitiveness Program of the National Research University Higher School of Economics.

[3]Supported by the Institute and Faculty of Actuaries, London.

[4]Supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2015R1A2A1A05001753).

MSC2010 subject classifications. Primary 62G07; secondary 62G20.

Key words and phrases. Density estimation, kernel smoothing, backfitting, chain Ladder.

one-dimensional components while estimating the density on S_2. The strategy may be put into practice by structuring the density in such a way that all components of the structured density are estimable with the observations in S_1. With this strategy, forecasting can be performed without extrapolation of parameters. This is likely to lead to more robust forecasting, because extrapolated parameters are often volatile. For time series extrapolation in particular, see Lee and Carter (1992), for example.

Lee et al. (2015) and Mammen, Martínez Miranda and Nielsen (2015) considered the perhaps simplest possible in-sample forecaster, where the joint density p has a multiplicative structure $p(x, y) = f_1(x) f_2(y)$ for some unknown univariate functions f_j with $S_1 = \{(x, y) : x \geq 0, y \geq 0, x + y \leq t_0\}$. In this setting, p is the joint density of two random variables X and Y, where X represents the start of something and Y is the development to some event from this starting point. These variables are observed only if the event occurs by a fixed calendar time t_0. Thus, $f_1(x)$ measures how many individuals are exposed or under risk and $f_2(y)$ represents duration or survival. The multiplicative form means that survival or duration has the same distribution independent of X. As was pointed out in Lee et al. (2015) and Mammen, Martínez Miranda and Nielsen (2015), this is a continuous type in-sample forecaster that extends classical actuarial and mortality forecasting methodologies based on multiplicative Poisson models being used every day in virtually all nonlife insurance companies around the world. In a nonparametric universe, the estimators resulting from the multiplicative Poisson models are structured histograms. Martínez-Miranda et al. (2013) showed the link between actuarial parametric chain ladder-type models [Kuang, Nielsen and Nielsen (2009)] and structured smoothing as considered in this paper.

The multiplicative structure $f_1(x) f_2(y)$ may be too simple for many settings. Nevertheless, the multiplicative model can be used as a baseline for more sophisticated models that deviate from this simple structure. This paper illustrates how powerful in-sample forecasting is when formulating, interpreting and analysing extensions of the simple multiplicative model. Actuaries have long tried to introduce the concept of operational time in the claims reserving modelling. The phrase "operational time" is taken from the literature of Poisson processes. When transforming the time axis with its operational time, an inhomogeneous Poisson process is transformed to a homogenous one; see Mikosch (2009) among many others. In the claims modelling framework, actuaries have been concerned about adjusting for changes in the speed of claims finalization over time. Many actuarial conference proceeding papers have been devoted to this topic and still are to this day. However, operational time or speed of claims finalization only had a short blossoming in the more formal academic actuarial literature; see Reid (1978), Taylor (1981, 1982) and Zehnwirth (1982). We believe that the topic of operational time died out in the actuarial literature, not because of lack of relevance, but because the mathematical challenges of formulating and analysing it became too overwhelming.

This paper introduces operational time to a general class of multiplicative models including actuarial, demographic and labour market applications taking advantage of the general in-sample forecasting formulation. We refer to Lee et al. (2015), Mammen, Martínez Miranda and Nielsen (2015) and Wilke (2016) for practical illustrations of multiplicative In-sample forecasting in actuarial science, demographics and the labour market. An alternative to operational time could be to add a calendar effect to the multiplicative model. While calendar effects are popular to talk about in actuarial science, they cause a number of difficulties, the most serious being the identifiability issue that some arbitrary linear trends can be added to or subtracted from the underlying model without changing the underlying model; see Kuang, Nielsen and Nielsen (2008a, 2008b, 2011). While the latter of these three papers does suggest practical implementation of identified forecasting procedures using calendar effects, there is still considerable uncertainty on how to forecast calendar effects in practice in the simple multiplicative forecasting model. Our Operational Time In-Sample Forecaster does not have any of these practical problems. It is immediate to construct a practical forecaster based on the operational time extension of the simple multiplicative In-Sample Forecaster.

In this paper, we consider a transformation, say ϕ, and a density model given by $p(x, y) = f_1(x) f_2(y\phi(x))$ on S_1. Our method and theory apply to a general type of support set S_1. The Operational Time In-Sample Forecaster can be understood as a structured model formulated in a density framework rather than in the regression framework considered in Mammen and Nielsen (2003). The model is formulated via a known smooth function of three one-dimensional unknown functions ϕ, f_1 and f_2. The estimation of ϕ, as discussed in Section 3, involves the estimation of the partial derivatives of the two-dimensional joint density function p by kernel smoothing. A naive application of the standard theory of kernel smoothing to the problem renders only a sub-optimal rate of convergence for the estimator of ϕ. Based on an innovative asymptotic analysis, we show that our estimator of ϕ achieves the optimal one-dimensional rate. Using this result, we also establish that the component functions f_j can be estimated with the optimal univariate rate.

There is a close relation between the multiplicative density model and the additive regression model. Thus, our approach may be extended to fundamental structured regression models studied in Jiang, Fan and Fan (2010), Yu, Park and Mammen (2008), Lee, Mammen and Park (2010, 2012), Zhang, Park and Wang (2013) among others. The multiplicative model with operational time corresponds to nonparametric regression models of the form $Z = m_1(X) + m_2(Y\phi(X)) + \varepsilon$ or $Z = m_1(X) + m_2(Y + \phi(X)) + \varepsilon$. The latter model is related to the nonparametric neural network models studied in Horowitz and Mammen (2007); see also recent work on composite function models by Juditsky, Lepski and Tsybakov (2009) and Baraud and Birgé (2014).

In-sample forecasting may be considered to be related to problems in survival analysis. In contrast with survival analysis, in-sample forecasting does not require full follow up of exposure and events, but is based only on the events that actually

happened and on a retrospective observation of the onset of these events. Therefore, there needs to be a lot less data to keep track of. For example, in-sample forecasting requires only keeping track of actual deaths of AIDS and retrospectively observed onset of AIDS, while most of survival analysis techniques need full follow up of how many individuals are under risk at any time (exposure), in addition to actual deaths of AIDS. The reason that in-sample forecasting needs fewer data requirements is that it estimates from data the equivalent of exposure in survival analysis. Our model is in some way related to accelerated failure time models. If one assumes that exposure is fully known and that one has only the components f_2 and ϕ in the model, then our model compares to an accelerated failure time model with X being a covariate; see Example VII 6.3 in Andersen et al. (1993). However, there are some differences. First of all, our approach is fully nonparametric. Second, our data are right truncated. Thus, exposure is not observed and it is only indirectly represented in our model via the component f_1 and estimated from the data. We therefore note that survival analysis techniques are not directly applicable in our model or in the application discussed in Section 7 in particular.

2. The model. We observe a random sample $\{(X_i, Y_i) : 1 \le i \le n\}$ from a density p supported on a subset $\mathcal{I}$ of the unit rectangle $[0, 1]^2$. The density $p(x, y)$ of (X_i, Y_i) is a multiplicative function

$$p(x, y) = f_1(x) f_2\big(y\phi(x)\big), \qquad (x, y) \in \mathcal{I}, \tag{2.1}$$

where f_1, f_2 and ϕ are unknown nonnegative functions bounded away from zero on their supports. We assume that f_1 and ϕ are supported on $[0, 1]$. We begin by considering the triangular support set $\mathcal{I} = \{(x, y) : 0 \le x, y \le 1, x + y \le 1\}$ of a rectangle $[0, 1]^2$ since the main idea of our approach can be best conveyed through the simple case. We discuss the method and theory for a general type of support set in Section 5.

In model (2.1), if x indicates the beginning of some development and y is the time this development takes, then $\phi(x)$ indicates a time transformation depending on the beginning of the development. When $\phi(x)$ gets bigger (smaller), time runs faster (slower) for the development beginning at x. From ad hoc analyses of practical applications of the In-Sample Forecaster "Double Chain Ladder" [Martínez Miranda, Nielsen and Verrall (2012)] to one of UK's largest global nonlife insurers, it has become clear that speed of time was increasing for almost every single dataset considered. This was the case for both the frequencies (number of claims) and the severities (size of claims). In the practical application of our model presented in Section 7 where frequencies from another nonlife insurer are considered, it can be concluded from our new operational time model, that speed of time also here is increasing. One likely explanation is of course that administration time, communication and reporting go faster as technology develops.

We start with the identification of the function ϕ in the model (2.1). The idea is also used for nonparametric estimation of the function, which we detail in the next section. Note that

$$\frac{\partial}{\partial x}\log p(x,y) = \frac{f_1'(x)}{f_1(x)} + \frac{f_2'(y\phi(x))}{f_2(y\phi(x))}y\phi'(x),$$
$$\frac{\partial}{\partial y}\log p(x,y) = \frac{f_2'(y\phi(x))}{f_2(y\phi(x))}\phi(x).$$

To represent ϕ in terms of the two partial derivatives, we think of a suitable contrast function $w(\cdot;x):\mathbb{R}\to\mathbb{R}$ for each $x\in[0,1)$, having the property that $\int_0^{1-x} w(y;x)\,dy = 0$. Then we have

$$\int_0^{1-x}\left(\frac{\partial}{\partial x}\log p(x,y)\right)w(y;x)\,dy = A(x)\phi'(x),$$
$$\int_0^{1-x}\left(\frac{\partial}{\partial y}\log p(x,y)\right)yw(y;x)\,dy = A(x)\phi(x),$$

where

$$A(x) = \int_0^{1-x}\frac{f_2'(y\phi(x))}{f_2(y\phi(x))}yw(y;x)\,dy.$$

If $A(x)\neq 0$ for all $x\in[0,1)$, then we get

$$\frac{\phi'(x)}{\phi(x)} = \frac{\int_0^{1-x}(\frac{\partial}{\partial x}\log p(x,y))w(y;x)\,dy}{\int_0^{1-x}(\frac{\partial}{\partial y}\log p(x,y))yw(y;x)\,dy}.$$

For the contrast function w, we take

$$w(y;x) = y\frac{\partial}{\partial y}\log p(x,y) - \frac{1}{1-x}\int_0^{1-x} y\frac{\partial}{\partial y}\log p(x,y)\,dy.$$

Note that $y\partial\log p(x,y)/\partial y = y\phi(x)f_2'(y\phi(x))/f_2(y\phi(x))$ is actually a function of $y\phi(x)$, and that with the choice of w we get

$$\frac{1}{1-x}\int_0^{1-x}\left(\frac{\partial}{\partial y}\log p(x,y)\right)yw(y;x)\,dy$$
$$= \frac{1}{\tau(x)}\int_0^{\tau(x)}\left(z\cdot\frac{f_2'(z)}{f_2(z)}\right)^2 dz - \left(\frac{1}{\tau(x)}\int_0^{\tau(x)} z\cdot\frac{f_2'(z)}{f_2(z)}\,dz\right)^2,$$

where $\tau(x) = (1-x)\phi(x)$. Thus, $A(x)>0$ if $zf_2'(z)/f_2(z)$ is not a function that is constant a.e. on $(0,\tau(x))$. Now, for x_0 fixed,

$$\ln\bigl(\phi(x)/\phi(x_0)\bigr) = \int_{x_0}^{x}\frac{\phi'(u)}{\phi(u)}\,du = \int_{x_0}^{x}\left[\frac{\int_0^{1-u}(\frac{\partial}{\partial u}\log p(u,y))w(y;u)\,dy}{\int_0^{1-u}(\frac{\partial}{\partial y}\log p(u,y))yw(y;u)\,dy}\right]du.$$

We choose $x_0 = 0$ and take the normalization $\phi(0) = 1$. For $j, k = 0, 1, 2$, define

$$G_{jk}(x) = \frac{1}{1-x}\int_0^{1-x}\left(\frac{\partial}{\partial x}\log p(x, y)\right)^j\left(y\frac{\partial}{\partial y}\log p(x, y)\right)^k dy. \tag{2.2}$$

Then we get

$$\phi(x) = \exp\left[\int_0^x \frac{G_{11}(u) - G_{10}(u)G_{01}(u)}{G_{02}(u) - G_{01}(u)^2}\,du\right]. \tag{2.3}$$

To assure that $A(x) > 0$ for any $x \in [0, 1)$, we make the following assumption:

(A1) For any small $c > 0$, $zf_2'(z)/f_2(z)$ is not a function that is constant a.e. on $(0, c)$.

Note that assumption (A1) concerns the behavior of the function $zf_2'(z)/f(z)$ near $z = 0$ only, since a function is not constant on $(0, c_1)$ if the function is not on $(0, c_2)$ for $c_2 < c_1$. The assumption is implied by the simpler one that there exists a small $c_0 > 0$ such that $zf_2'(z)/f(z)$ is strictly monotone on $(0, c_0)$, or that its derivative is not zero at $z = 0$ in case it is continuously differentiable.

Next, we discuss the identifiability of the component functions f_1 and f_2. The following arguments are based on the identifiability of ϕ, which we have just proved. The two component functions f_1 and f_2 are identifiable only up to a multiplicative constant. Hence, we put the constraint on the first component that

$$\int_0^1 f_1(x)\,dx = 1. \tag{2.4}$$

Let $\mu_1(x) = \log f_1(x)$ and $\mu_2(z) = \log f_2(z)$. Suppose that $\mu_1(x) + \mu_2(y\phi(x)) = 0$ for all $(x, y) \in \mathcal{I}$. By differentiating both sides with respect to y, we get

$$\phi(x)\mu_2'\big(y\phi(x)\big) = 0.$$

Since we assume that $\phi(x) > 0$ for all $x \in [0, 1]$, this implies $\mu_2'(y\phi(x)) = 0$ for all $(x, y) \in \mathcal{I}$. Thus, μ_2 is constant on its domain, so is μ_1. Due to the constraint (2.4), we have $\mu_1 \equiv 0$ on $[0, 1]$ so that $\mu_2 \equiv 0$ on its domain as well.

THEOREM 1. *Assume that the two component functions f_j and the time transformation ϕ in the model* (2.1) *are differentiable, nonnegative and bounded away from zero on their supports. Assume also that* (A1) *holds. Then the three functions ϕ, f_1 and f_2 are identifiable under the constraint* (2.4).

3. Estimation of time transformation. Here, we describe the estimation of the time transformation ϕ based on the local quadratic smoothing technique. Note that Lee et al. (2015) suggested to use the local linear smoothing method since their model involves only the estimation of the joint density function. Here, ϕ is identified through the partial derivatives of the joint density p, as is seen from (2.2) and (2.3). Therefore, one may want to use local quadratic smoothing to ensure stable

performance at the boundary area of $\mathcal{I}$ in the estimation of the partial derivatives. Indeed, in our preliminary simulation study we found that local linear smoothing produced quite bad estimates of the first partial derivatives.

To define the estimator of ϕ based on local quadratic smoothing, let

$$\mathbf{a}(u,v;x,y) = \big(1,(u-x)/h_1,(v-y)/h_2,(u-x)^2/h_1^2,$$
$$(u-x)(v-y)/h_1h_2,(v-y)^2/h_2^2\big)^\top,$$
$$\mathbf{A}(x,y) = \int_{\mathcal{I}} \mathbf{a}(u,v;x,y)\mathbf{a}(u,v;x,y)^\top h_1^{-1}h_2^{-1}K\left(\frac{u-x}{h_1}\right)K\left(\frac{v-y}{h_2}\right)du\,dv,$$

where (h_1,h_2) is the bandwidth vector and K is a symmetric univariate probability density function. Also, define

$$\hat{\mathbf{b}}(x,y) = n^{-1}\sum_{i=1}^{n}\mathbf{a}(X_i,Y_i;x,y)h_1^{-1}h_2^{-1}K\left(\frac{X_i-x}{h_1}\right)K\left(\frac{Y_i-y}{h_2}\right).$$

The local quadratic density estimators of $p(x,y)$, $\frac{\partial}{\partial x}p(x,y)$ and $\frac{\partial}{\partial y}p(x,y)$, respectively, are then defined by $\hat{\eta}_{00}(x,y)$, $\hat{\eta}_{10}(x,y)/h_1$ and $\hat{\eta}_{01}(x,y)/h_2$, respectively, where

$$(\hat{\eta}_{00},\hat{\eta}_{10},\hat{\eta}_{01},\hat{\eta}_{20},\hat{\eta}_{11},\hat{\eta}_{02})^\top = \mathbf{A}^{-1}\hat{\mathbf{b}}. \tag{3.1}$$

The above estimators of the joint density p and its partial derivatives are similar in spirit to the local linear density estimators studied in Cheng (1997). Putting these into formula (2.2), we get the estimators $\hat{G}_{jk}(x)$ of $G_{jk}(x)$, and thus the estimator of ϕ defined by

$$\hat{\phi}(x) = \exp\left[\int_0^x \frac{\hat{G}_{11}(u)-\hat{G}_{10}(u)\hat{G}_{01}(u)}{\hat{G}_{02}(u)-\hat{G}_{01}(u)^2}\,du\right]. \tag{3.2}$$

The convergence rate of the estimator $\hat{\phi}$ depends on those of the estimators $\hat{\eta}_{jk}$ of the joint density and its partial derivatives. For simplicity of presentation we write

$$p_{jk}(x,y) = \frac{\partial^{j+k}}{\partial x^j\,\partial y^k}p(x,y).$$

If p is twice partially continuously differentiable, then from an expansion of $p(u,v)$ for (u,v) around (x,y) one gets that $E\hat{\eta}_{jk}(x,y) - h_1^j h_2^k p_{jk}(x,y) = o(h_1^2+h_2^2)$ for (j,k) with $0\le j,k\le 1$ and $j+k\le 1$. Furthermore, one has $\hat{\eta}_{jk}(x,y) - E\hat{\eta}_{jk}(x,y) = O_p(n^{-1/2}h_1^{-1/2}h_2^{-1/2})$; see Ruppert and Wand (1994) or Fan, Heckman and Wand (1995) among others. These imply that the estimators of the first-order partial derivatives have the convergence rate $O_p(n^{-1/2}h_1^{-3/2}h_2^{-1/2})+o_p(h_1+h_1^{-1}h_2^2)$ or $O_p(n^{-1/2}h_1^{-1/2}h_2^{-3/2})+o_p(h_1^2h_2^{-1}+$

$h_2)$. Note that the estimator $\hat{\phi}(x)$ involves two integrations of the estimators of the first-order partial derivatives, one for each coordinate; see the definitions (2.2) and (3.2). In the standard kernel smoothing theory, it is well known that a nontrivial integration of a kernel estimator makes the stochastic part get smaller by an order of $h^{1/2}$, where h is the size of the bandwidth that is used for local smoothing along the line of the integration. This is mainly because the "local average" turns into a "global average" along the lines of the integration; see Mammen, Park and Schienle (2014), for example. For the bias term, an integration does not reduce the order of magnitude in general, however. A direct application of this standard theory to $\hat{\phi}(x)$ would give the rate $O_p(n^{-1/2}\min\{h_1,h_2\}^{-1}) + o_p(\max\{h_1,h_2\})$. One may improve the rate for the bias part to $O_p(\max\{h_1,h_2\}^2)$ if one assumes three times partial differentiability, which would lead to the two-dimensional rate $n^{-1/3}$ at best by choosing $h_1 \sim h_2 \sim n^{-1/6}$.

In the theorem below, however, we show that our estimator $\hat{\phi}$ achieves the univariate rate of convergence $n^{-2/5}$ under the condition that p is twice partially continuously differentiable. Before we state the theorem, here we give an intuition behind and heuristic argument for the surprising results. Let $m_2(u) = \log f_2(e^u)$ and $m_3(u) = \log\phi(u)$, where f_2 is the second component function in our model (2.1). For an arbitrarily small constant $\epsilon > 0$, define a bivariate function F_ϵ by

$$F_\epsilon(x,t) = \int_t^{t+\epsilon} \log p\big(x,e^z\big)\,dz - \int_{t-\epsilon}^{t} \log p\big(x,e^z\big)\,dz \tag{3.3}$$

on $\{(x,t): 0 \le x < 1, t \le \log(1-x) - \epsilon\}$. Then F_ϵ may be expressed in terms of a univariate function and ϕ. Indeed, letting $H_\epsilon(t) = \int_0^\epsilon [m_2(z+t) - m_2(z-\epsilon+t)]\,dz$, we get

$$\begin{aligned} F_\epsilon(x,t) &= \int_0^\epsilon \big[m_2\big(z+t+m_3(x)\big) - m_2\big(z-\epsilon+t+m_3(x)\big)\big]\,dz \\ &= H_\epsilon\big(t+m_3(x)\big). \end{aligned}$$

Recall our normalization $\phi(0) = 1$ for ϕ, so that $m_3(0) = 0$. Thus, for $t \le \log\tau(x) - \epsilon$ we get

$$F_\epsilon\big(x, -m_3(x)+t\big) = H_\epsilon(t) = F_\epsilon(0,t). \tag{3.4}$$

From the definition of F_ϵ at (3.3) we note that F_ϵ may be estimated with the univariate rate, because of the integration. Now, due to (3.4) one may identify m_3, thus ϕ, by identifying F_ϵ, provided that, for any $x \in [0,1)$, one finds $t_0 < \log\tau(x) - \epsilon$ such that $\partial F_\epsilon(x,t)/\partial t$ is not zero at $t = -m_3(x) + t_0$. The latter also means that m_3 can be estimated with the same accuracy as F_ϵ. The condition on $\partial F_\epsilon(x,t)/\partial t$ is implied by the assumption (A1). To see this, we note that

$$\frac{\partial}{\partial t} F_\epsilon(x,t)\bigg|_{t=-m_3(x)+t_0} = H_\epsilon'(t_0) = m_2(t_0+\epsilon) - 2m_2(t_0) + m_2(t_0-\epsilon).$$

Observing that $m_2'(t) = e^t f_2'(e^t)/f_2(e^t)$, the assumption (A1) is equivalent to the condition that, for any $C > 0$, m_2' is not a function that is constant a.e. on $(-\infty, -C)$. Thus, (A1) implies that, for any $C > 0$, there exists $t_0 < -C - \epsilon$ such that $H_\epsilon'(t_0) \neq 0$.

We now state our theorem for the rate of $\hat{\phi}(x)$. Below, we give a pointwise convergence rate for $x \in [0, 1)$, excluding the point $x = 1$. Also, we present the rates for the integrated squared and the uniform errors on an interval $[0, 1 - \epsilon]$ for an arbitrarily small $\epsilon > 0$. The reason we exclude the point $x = 1$ is that the marginal density of X vanishes at $x = 1$ even though the joint density f is bounded away from zero on its support. This is due to the triangular shape of the support set $\mathcal{I}$. Thus, the consistent estimation of $\phi(x)$ as x approaches to the end point 1 is not possible. We make the following additional assumptions:

(A2) The joint density function p is twice partially continuously differentiable and bounded away from zero;
(A3) The kernel K is supported on $[-1, 1]$, symmetric and Lipschitz continuous;
(A4) The bandwidths h_1 and h_2 are of order $n^{-1/5}$.

THEOREM 2. *Assume that the conditions of Theorem 1 and conditions* (A2)–(A4) *are satisfied. Then we get for* $x \in [0, 1)$ *that*

$$\hat{\phi}(x) - \phi(x) = O_p\big(n^{-2/5}\big). \tag{3.5}$$

Furthermore, for an arbitrarily small $\epsilon > 0$, *it holds that*

$$\int_0^{1-\epsilon} \big(\hat{\phi}(x) - \phi(x)\big)^2 \, dx = O_p\big(n^{-4/5}\big), \tag{3.6}$$

$$\sup_{x \in [0, 1-\epsilon]} |\hat{\phi}(x) - \phi(x)| = O_p\big(n^{-2/5}\sqrt{\log n}\big). \tag{3.7}$$

In the proof of Theorem 2 given in the Appendix, one sees that $\hat{\phi}(x)$ is not a local smoother. If one looks at the term $J_1(x)$ that is discussed at the end of the proof, one finds that this quadratic form is of order $O_p(n^{-2/5})$ and not negligible in the first order. By definition all observations (X_i, Y_i) enter $J_1(x)$ with weights of the same magnitude. Thus, this term does not rely only on local information. The same holds for $\hat{\phi}(x)$. It is calculated using all observations, not only those (X_i, Y_i) with X_i in a shrinking neighborhood of x. This makes $\hat{\phi}$ quite different from a kernel smoother.

4. Estimation of component functions. Suppose we know the true time transformation ϕ. Then we would convert the dataset (X_i, Y_i) to (X_i, Z_i) with $Z_i = Y_i \phi(X_i)$, and estimate the component functions f_1 and f_2 from the converted dataset. The density function of (X_i, Z_i) equals $p(x, z/\phi(x))/\phi(x)$, and

the model (2.1) reduces to

$$p\left(x, \frac{z}{\phi(x)}\right) = f_1(x) f_2(z). \tag{4.1}$$

Recall that we take the normalization $\phi(0) = 1$. This means that time runs as real time at the starting point, and that the set $\{(u, v\phi(u)) : (u, v) \in \mathcal{I}\}$ includes the two edge points $(0, 1)$ and $(1, 0)$ of the triangle $\mathcal{I}$.

For the estimation of the component functions f_1 and f_2 at points x and z, respectively, we need sufficient data X_i and Z_i around x and z. We estimate f_1 and f_2 on intervals where the marginal densities of X_i and Z_i, respectively, are bounded away from zero. Note that the marginal density of X_i at x and that of Z_i at z are given by $\int_0^{(1-x)\phi(x)} p(x, z/\phi(x))/\phi(x)\,dz$ and $\int_{x:\tau(x)\geq z} p(x, z/\phi(x))/\phi(x)\,dx$, respectively, where τ defined by $\tau(x) = (1 - x)\phi(x)$. We assume:

(A5) τ is strictly decreasing.

Condition (A5) simplifies the description of the method and the presentation of its theory. In this case $\{x \in [0, 1] : \tau(x) \geq z\} = [0, \tau^{-1}(z)]$, and $\tau^{-1}(z) = 0$ holds only for $z = 1$. The method we describe below and its theory are based on this condition on τ. We discuss a general case at the end of this section.

The marginal density function of X_i equals zero at $x = 1$ and that of Z_i is zero at $z = 1$, even if the joint density p is bounded away from zero on its support. For a set of (x, z) where we estimate the component functions f_1 and f_2, we take

$$\begin{aligned} I &\equiv \big\{\big(u, v\phi(u)\big) : u \leq 1 - \epsilon, v\phi(u) \leq 1 - \epsilon, (u, v) \in \mathcal{I}\big\} \\ &= \big\{(u, w) : 0 \leq u \leq 1 - \epsilon, 0 \leq w \leq (1 - \epsilon) \wedge \tau(u)\big\} \end{aligned} \tag{4.2}$$

for an arbitrarily small $\epsilon > 0$. The projections of the set I onto x- and z-axis equal $[0, 1 - \epsilon]$. Thus, we estimate both f_1 and f_2 on an interval $[0, 1 - \epsilon]$. Define $I_1(z) = \{x : (x, z) \in I\}$ and $I_2(x) = \{z : (x, z) \in I\}$. Note that

$$I_1(z) = \big[0, (1 - \epsilon) \wedge \tau^{-1}(z)\big], \qquad I_2(x) = \big[0, (1 - \epsilon) \wedge (1 - x)\phi(x)\big].$$

Furthermore,

$$\inf_{z\in[0,1-\epsilon]} \operatorname{mes}\big(I_1(z)\big) > 0, \qquad \inf_{x\in[0,1-\epsilon]} \operatorname{mes}\big(I_2(x)\big) > 0,$$

where $\operatorname{mes}(A)$ denotes the Lebesgue measure of a set A. It follows that the marginalization of $p(x, z/\phi(x))$ along $I_1(z)$ and the one along $I_2(x)$ are bounded away from zero for $z \in [0, 1 - \epsilon]$ and $x \in [0, 1 - \epsilon]$, respectively, that is,

$$\begin{aligned} &\inf_{z\in[0,1-\epsilon]} \int_{I_1(z)} p\big(x, z/\phi(x)\big)\,dx > 0, \\ &\inf_{x\in[0,1-\epsilon]} \int_{I_2(x)} p\big(x, z/\phi(x)\big)\,dz > 0, \end{aligned} \tag{4.3}$$

provided that p is bounded away from zero on $\mathcal{I}$.

We take the marginalization technique of Lee et al. (2015) to estimate the component functions. For now, we assume the true ϕ is known. Integrating both sides of (4.1) along the lines $I_1(z)$ and $I_2(x)$ gives

$$\begin{aligned} f_1(x) &= \Big(\int_{I_2(x)} f_2(z)\,dz\Big)^{-1} \int_{I_2(x)} p\big(x, z/\phi(x)\big)\,dz, \\ f_2(z) &= \Big(\int_{I_1(z)} f_1(x)\,dx\Big)^{-1} \int_{I_1(z)} p\big(x, z/\phi(x)\big)\,dx. \end{aligned} \tag{4.4}$$

The inverses in (4.4) are well defined for all $x, z \in [1-\epsilon]$ due to (4.3). Set $\vartheta = \int_I p(x, z/\phi(x))/\phi(x)\,dx\,dz$. Then $\vartheta^{-1} f_1(x) f_2(z)\phi(x)^{-1}$ is a density on I. Let $\hat{p}$ be an estimator of the joint density p. Putting the constraint $\int_0^{1-\epsilon} f_1(x)\,dx = 1$ on the estimator of the first component f_1, our estimator of (f_1, f_2) is defined to be the solution $(\tilde{f}_1, \tilde{f}_2)$ of the system of equations

$$\begin{aligned} \tilde{f}_1(x) &= \tilde{\theta}_1 \Big(\int_{I_2(x)} \tilde{f}_2(z)\,dz\Big)^{-1} \int_{I_2(x)} \hat{p}\big(x, z/\phi(x)\big)\,dz, \\ \tilde{f}_2(z) &= \tilde{\theta}_2 \Big(\int_{I_1(z)} \tilde{f}_1(x)\,dx\Big)^{-1} \int_{I_1(z)} \hat{p}\big(x, z/\phi(x)\big)\,dx, \end{aligned} \tag{4.5}$$

where $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are chosen so that

$$\int_0^{1-\epsilon} \tilde{f}_1(x)\,dx = 1, \qquad \int_I \tilde{f}_1(x)\tilde{f}_2(z)/\phi(x)\,dx\,dz = \tilde{\vartheta}, \tag{4.6}$$

and $\tilde{\vartheta} = n^{-1}\sum_{i=1}^n I[X_i \le 1-\epsilon, Y_i\phi(X_i) \le 1-\epsilon]$.

Since ϕ, in the above construction of $\tilde{f}_1$ and $\tilde{f}_2$, is unknown, we replace it by the estimator $\hat{\phi}$ studied in Section 3. For this, we define a version of I for a general time transformation function φ by

$$I(\varphi) = \big\{(x, z) : 0 \le x \le 1-\epsilon, 0 \le z \le (1-\epsilon) \wedge \tau(x; \varphi)\big\}$$

with $\tau(x; \varphi) = (1-x)\varphi(x)$, and those versions of $I_1(z)$ and $I_2(x)$, respectively, by

$$I_1(z, \varphi) = \big\{x \in [0, 1-\epsilon] : \tau(x; \varphi) \ge z\big\}, \qquad I_2(x, \varphi) = \big[0, (1-\epsilon) \wedge \tau(x; \varphi)\big].$$

Then the estimators $\hat{f}_1$ and $\hat{f}_2$ of the components f_1 and f_2, respectively, solve the system of equations (4.5) subject to the constraints (4.6) with ϕ, I, $I_1(z)$, $I_2(x)$ and $\tilde{\vartheta}$ being replaced by $\hat{\phi}$, $I(\hat{\phi})$, $I_1(z, \hat{\phi})$, $I_2(x, \hat{\phi})$ and $\hat{\vartheta} = n^{-1}\sum_{i=1}^n I[X_i \le 1-\epsilon, Y_i\hat{\phi}(X_i) \le 1-\epsilon]$, respectively. We denote the constraining constants $\tilde{\theta}_j$ in (4.5) by $\hat{\theta}_j$ in this case.

For the estimator $\hat{p}$ of the joint density p in (4.5), we suggest to use the local linear estimator at this stage. This is because at this time we only need an estimator

of the joint density itself, not its derivatives. Specifically, we estimate p by $\hat{\xi}_{00}$, where $(\hat{\xi}_{00}, \hat{\xi}_{10}, \hat{\xi}_{01})^\top = \mathbf{C}^{-1}\hat{\mathbf{d}}$ with

$$\mathbf{C}(x,y) = \int_{\mathcal{I}} \mathbf{c}(u,v;x,y)\mathbf{c}(u,v;x,y)^\top g_1^{-1} g_2^{-1} K\left(\frac{u-x}{g_1}\right) K\left(\frac{v-y}{g_2}\right) du\, dv,$$

$$\hat{\mathbf{d}}(x,y) = n^{-1}\sum_{i=1}^{n} \mathbf{c}(X_i, Y_i; x, y) g_1^{-1} g_2^{-1} K\left(\frac{X_i - x}{g_1}\right) K\left(\frac{Y_i - y}{g_2}\right),$$

and $\mathbf{c}(u,v;x,y) = (1, (u-x)/g_1, (v-y)/g_2)^\top$. Here, the bandwidth pair (g_1, g_2) may be different from (h_1, h_2) in the estimation of ϕ.

According to Lee et al. (2015), the estimators $\tilde{f}_j$ that are based on the true time transformation ϕ have the following uniform convergence rate:

$$\sup_{u\in[0,1-\epsilon]} |\tilde{f}_j(u) - f_j(u)| = O_p\big(n^{-1/2}\min\{g_1, g_2\}^{-1/2}\sqrt{\log n} + g_1^2 + g_2^2\big).$$

Thus, if one takes $g_1 \sim g_2 \sim n^{-1/5}$, then one gets the univariate rate $O_p(n^{-2/5}\sqrt{\log n})$. Our primary interest is to assess the effect of estimating ϕ in the estimation of f_1 and f_2. The following theorem demonstrates that the estimation of ϕ contributes to $\hat{f}_j - f_j$ an additional term that is of the same order as the estimation error $\hat{\phi} - \phi$. To state the theorem, we think of a space of quadruples where a quadruple in the space have two constants and two univariate functions. Define a nonlinear operator $\mathcal{G}(\boldsymbol{\eta}, \mathbf{g}, \phi)$, which maps the space of quadruples $(\boldsymbol{\eta}, \mathbf{g})$ to itself, by $\mathcal{G}(\boldsymbol{\eta}, \mathbf{g}, \phi)_1 = 1 - \int_0^{1-\epsilon} f_1(x)(1 + g_1(x))\, dx$ and

$$\mathcal{G}(\boldsymbol{\eta}, \mathbf{g}, \phi)_2 = \vartheta - \int_I f_1(x) f_2(z)\big(1 + g_1(x)\big)\big(1 + g_2(z)\big)\frac{1}{\phi(x)}\, dz\, dx,$$

$$\mathcal{G}(\boldsymbol{\eta}, \mathbf{g}, \phi)_3(u) = \int_{I_2(u)} \big[(1+\eta_1)p\big(u, z/\phi(u)\big) - f_1(u) f_2(z)\big(1 + g_1(u)\big)\big(1 + g_2(z)\big)\big]\, dz,$$

$$\mathcal{G}(\boldsymbol{\eta}, \mathbf{g}, \phi)_4(u) = \int_{I_1(u)} \big[(1+\eta_2)\hat{p}\big(x, u/\phi(x)\big) - f_1(x) f_2(u)\big(1 + g_1(x)\big)\big(1 + g_2(u)\big)\big]\, dx.$$

Let $\mathcal{G}'(\mathbf{0}, \mathbf{0}, \phi)$ denote the Fréchet derivative of $\mathcal{G}(\cdot, \cdot, \phi)$ at $(\mathbf{0}, \mathbf{0})$. It is an invertible linear operator. Let $\tilde{\Delta}_1$ and $\tilde{\Delta}_2$ be the last two entries of $\mathcal{G}'(\mathbf{0}, \mathbf{0}, \phi)^{-1}\tilde{\boldsymbol{\delta}}$, where $\tilde{\boldsymbol{\delta}} = (0, \tilde{\delta}_2, \tilde{\delta}_3, \tilde{\delta}_4)^\top$ and

$$\tilde{\delta}_2 = -\int_I f_1(x)\big(f_2\big(z\phi(x)/\hat{\phi}(x)\big) - f_2(z)\big)/\phi(x)\, dz\, dx,$$

$$\tilde{\delta}_3(x) = -\int_{I_2(x)} f_1(x)\big(f_2\big(z\phi(x)/\hat{\phi}(x)\big) - f_2(z)\big)\, dz, \tag{4.7}$$

$$\tilde{\delta}_4(z) = -\int_{I_1(z)} f_1(x)\big(f_2\big(z\phi(x)/\hat{\phi}(x)\big) - f_2(z)\big)\, dx.$$

THEOREM 3. *Assume that conditions of Theorem 1 hold and that the conditions* (A2), (A3) *and* (A5) *are satisfied. Assume also that the bandwidths* g_j *satisfy* $g_j \to 0$ *and* $ng_1g_2/\log n \to \infty$. *If* $\sup_{x\in[0,1-\epsilon]} |\hat{\phi}(x) - \phi(x)| = O_p(\varepsilon_n)$ *and* $\sup_{(x,y)\in\mathcal{I}} |\hat{p}(x,y) - p(x,y)| = O_p(\varepsilon_n')$, *then* $\hat{f}_j(u) - f_j(u) = \tilde{f}_j(u) - f_j(u) + f_j(u)\tilde{\Delta}_j(u) + O_p(n^{-1/2} + \varepsilon_n^2 + \varepsilon_n'^2 + \varepsilon_n g_1^2 g_2^{-1} + \varepsilon_n g_2 + \varepsilon_n n^{-1/2} g_2^{-3/2}\sqrt{\log n}) + o_p(g_1^2 + g_2^2)$, *for each fixed* $u \in [0,1)$, *and also uniformly for* $u \in [0, 1-\epsilon]$ *for an arbitrarily small* $\epsilon > 0$.

According to Theorem 2, $\varepsilon_n = n^{-2/5}\sqrt{\log n}$. Also, one has $\tilde{\Delta}_j(u) = O_p(n^{-2/5})$ for each fixed $u \in [0,1)$, and $\tilde{\Delta}_j(u) = O_p(n^{-2/5}\sqrt{\log n})$ uniformly for $u \in [0, 1-\epsilon]$. According to Theorem 4 of Lee et al. (2015), one has $\tilde{f}_j(u) - f_j(u) = O_p(n^{-2/5})$ for each fixed $u \in [0,1)$, and $\tilde{f}_j(u) - f_j(u) = O_p(n^{-2/5}\sqrt{\log n})$ uniformly for $u \in [0, 1-\epsilon]$. If we take the bandwidths $g_1 \sim g_2 \sim n^{-1/5}$, then $\varepsilon_n' = n^{-3/10}\sqrt{\log n}$. From Theorem 3, we obtain the following corollary.

COROLLARY 1. *Assume the conditions of Theorems 2 and 3 hold. If* $g_1 \sim g_2 \sim n^{-1/5}$, *then* $\hat{f}_j(u) - f_j(u) = \tilde{f}_j(u) - f_j(u) + f_j(u)\tilde{\Delta}_j(u) + o_p(n^{-2/5})$ *for each fixed* $u \in [0,1)$, *and also uniformly for* $u \in [0, 1-\epsilon]$ *for an arbitrarily small* $\epsilon > 0$.

The above corollary demonstrates that our estimators of the component functions f_j achieve the optimal uniform rate $O_p(n^{-2/5}\sqrt{\log n})$ as well as the optimal pointwise rate $O_p(n^{-2/5})$ in one-dimensional smoothing, under the condition that the joint density is twice partially continuously differentiable.

As we mentioned earlier in this section, we describe our method of estimating f_j and prove Theorem 3 under the assumption that τ is strictly decreasing. In the general case without this assumption, the component function f_2 sits on the interval $[0, \max_{x\in[0,1-\epsilon]} \tau(x)]$, so that one may estimate f_2 in an interval $[0, \max_{x\in[0,1-\epsilon]} \tau(x) - \epsilon]$ for an arbitrarily small $\epsilon > 0$. In this case, the set that corresponds to $I_1(u)$ will be a union of several intervals for some points u, and the procedure may be described along the lines of our presentation, but with more involved notation. The conclusion of Theorem 3 is also valid for $\hat{f}_1$ in the general case. For $\hat{f}_2$, it remains to hold uniformly for u in the interval $[0, \max_{x\in[0,1-\epsilon]} \tau(x) - \epsilon]$ with arbitrarily small neighborhoods of those points $u = \tau(x)$ for x with $\tau'(x) = 0$, being excluded. This can be seen from the fact that, in our proof of the theorem given in the supplement [Lee et al. (2016)], we use the condition $\tau' \neq 0$ only for

$$\operatorname{mes}\bigl(I_1(z,\hat{\phi}) \triangle I_1(z,\phi)\bigr) = O_p(\varepsilon_n).$$

To give more insight into how the theory depends on the shape of the function τ, we note that the second component function f_2 is identified by the marginalization

over the set $I_1(u) \equiv I_1(u;\phi)$, see the second equation at (4.4). This means that the accuracy of estimating f_2 depends on that of estimating ϕ through the difference between the lengths of the sets $I_1(u)$ and $I_1(u;\hat{\phi})$. If u is a point such that $u = \tau(x_0)$ for some x_0 with $\tau'(x_0) = 0$, then the estimation error of $\hat{\phi}$ is magnified in the difference between the two lengths. To see this, suppose that $\tau''(x) < 0$ and $\hat{\phi}(x) > \phi(x)$ for x in a neighborhood of x_0. Then, for a small constant $c > 0$, $I_1(u) \cap [x_0 - c, x_0 + c] = \{x \in [x_0 - c, x_0 + c] : \tau(x) \geq u\} = \{x_0\}$ since

$$\tau(x) \simeq u + \frac{1}{2}\tau''(x_0)(x - x_0)^2.$$

On the other hand, $I_1(u;\hat{\phi}) \cap [x_0 - c, x_0 + c] \supset [x_0 - d_n, x_0 + d_n]$, where $d_n = (\text{constant}) \times (\inf_{x\in[x_0-c,x_0+c]} |\hat{\phi}(x) - \phi(x)|)^{1/2}$ since

$$\tau(x,\hat{\phi}) \simeq \tau(x,\hat{\phi}) - \tau(x,\phi) + u + \frac{1}{2}\tau''(x_0)(x - x_0)^2.$$

From the above discussion, we see that the remainder term in the uniform expansion of $\hat{f}_2 - f_2$ over the whole interval $[0, 1 - \varepsilon]$ in Theorem 3 has $O_p(\varepsilon_n)$ instead of $O_p(\varepsilon_n^2)$, provided that $\tau''(x) \neq 0$ for all x in $(0, 1)$.

5. Extension to general support set. In this section, we extend the method and theory to a general type of support set $\mathcal{I}$ where the data (X_i, Y_i) are observed. Without of loss of generality, we assume that the projections of the support set $\mathcal{I}$ onto the x- and y-axis equal $[0, 1]$. For each $x \in [0, 1]$, define $\mathcal{I}_2(x) = \{y \in [0, 1] : (x, y) \in \mathcal{I}\}$. In the case of the triangular support that we considered in Sections 2, 3 and 4, $\mathcal{I}_2(x) = [0, 1 - x]$. Define

$$\mathcal{I}_1 = \{x \in [0, 1] : \text{mes}(\mathcal{I}_2(x)) \neq 0\}.$$

Then we get (2.3) for $x \in \mathcal{I}_1$ with $G_{jk}(x)$ now being defined by

$$G_{jk}(x) = \frac{1}{\text{mes}(\mathcal{I}_2(x))} \int_{\mathcal{I}_2(x)} \left(\frac{\partial}{\partial x} \log p(x, y)\right)^j \left(y\frac{\partial}{\partial y} \log p(x, y)\right)^k dy.$$

Condition (A1), for the identifiability of ϕ, f_1 and f_2, is now generalized to:

(A1′) For all $x \in \mathcal{I}_1$, $zf_2'(z)/f_2(z)$ is not a function that is constant a.e. on $\{y\phi(x) : y \in \mathcal{I}_2(x)\}$.

We obtain the following analogue of Theorem 1 for the general support set $\mathcal{I}$.

THEOREM 4. *Assume that the two component functions f_j and the time transformation ϕ in the model* (2.1) *are differentiable, nonnegative and bounded away from zero on their supports. Assume also that* (A1′) *holds and that the set $\mathcal{I}_1$ is dense on* $[0, 1]$. *Then the three functions ϕ, f_1 and f_2 are identifiable under the constraint* (2.4).

The estimation of ϕ is defined as (3.2) with $\hat{G}_{jk}$ now being redefined by

$$\hat{G}_{jk}(x) = \frac{1}{\operatorname{mes}(\mathcal{I}_2(x))} \int_{\mathcal{I}_2(x)} \left(\frac{\hat{\eta}_{10}(x, y)/h_1}{\hat{\eta}_{00}(x, y)} \right)^j \left(y \frac{\hat{\eta}_{01}(x, y)/h_2}{\hat{\eta}_{00}(x, y)} \right)^k dy.$$

Note that one can estimate $\phi(x)$ only for x with $\operatorname{mes}(\mathcal{I}_2(x)) > 0$. This was the reason we exclude the point $x = 1$ for the estimation of ϕ in the case of the triangular support. In the general case we consider here, we exclude the point $x \notin \mathcal{I}_1$. Also, to get the L_2 and the uniform convergence results as in Theorem 2, we consider the set $\mathcal{I}_{1,\epsilon} = \{x : \operatorname{mes}(\mathcal{I}_2(x)) \geq \epsilon\}$ for an arbitrarily small $\epsilon > 0$.

THEOREM 5. *Assume that the conditions of Theorem 4 and the conditions* (A2)–(A4) *are satisfied. Then we get for $x \in \mathcal{I}_1$ that $\hat{\phi}(x) - \phi(x) = O_p(n^{-2/5})$. Furthermore, for an arbitrarily small $\epsilon > 0$, it holds that*

$$\int_{\mathcal{I}_{1,\epsilon}} \big(\hat{\phi}(x) - \phi(x)\big)^2 dx = O_p\big(n^{-4/5}\big),$$

$$\sup_{x \in \mathcal{I}_{1,\epsilon}} |\hat{\phi}(x) - \phi(x)| = O_p\big(n^{-2/5}\sqrt{\log n}\big).$$

Now we extend the method of estimating the component functions f_j to the general support set. As in the case of the triangular support, one may estimate f_1 and f_2, respectively, only on the sets where the marginal densities of X and Z are strictly positive. We find a version of the set I defined at (4.2). For a subset S of the support $\{(x, y\phi(x)) : (x, y) \in \mathcal{I}\}$ of the joint density of (X, Z), let

$$I_1(z; S) = \big\{x : (x, z) \in S\big\}, \qquad I_2(x; S) = \big\{z : (x, z) \in S\big\}.$$

Taking a small $\delta > 0$ we choose I, the set where we estimate f_j, to be the largest subset S such that

$$\operatorname{mes}\big(I_1(z, S)\big) \geq \delta, \qquad \operatorname{mes}\big(I_2(x, S)\big) \geq \delta$$

for all x and z in the projections of S onto the x- and z-axis, respectively. We write $I_1(z) = I_1(z, I)$ and $I_2(x) = I_2(x, I)$ for simplicity. We estimate f_j on the set I_j, where

$$I_1 = \big\{x : \big(x, y\phi(x)\big) \in I \text{ for some } y \in [0, 1]\big\},$$
$$I_2 = \big\{z : (x, z) \in I \text{ for some } x \in [0, 1]\big\}.$$

With these modified definitions of the sets $I_1(z)$ and $I_2(x)$, the estimators of f_j based on the true ϕ may be defined as at (4.4) and (4.5), now with the constraints

$$\int_{I_1} \tilde{f}_1(x)\, dx = 1, \qquad \int_I \tilde{f}_1(x)\, \tilde{f}_2(z)/\phi(x)\, dx\, dz = \tilde{\vartheta},$$

where $\tilde{\vartheta}$ is redefined as $\tilde{\vartheta} = n^{-1}\sum_{i=1}^{n} I[(X_i, Y_i\phi(X_i)) \in I]$. The estimators $\hat{f}_j$ based on $\hat{\phi}$ is then obtained by simply replacing I, $I_1(z)$, $I_2(x)$ and $\tilde{\vartheta}$ in the definition of $\tilde{f}_j$ by $I(\hat{\phi})$, $I_1(z, \hat{\phi})$, $I_2(z, \hat{\phi})$ and $\hat{\vartheta}$, respectively, where $I(\varphi)$, $I_1(z, \varphi)$ and $I_2(x, \varphi)$ for a general time transformation φ are defined as I, $I_1(z)$ and $I_2(x)$ with ϕ being replaced by φ, and $\hat{\vartheta} = n^{-1}\sum_{i=1}^{n} I[(X_i, Y_i\hat{\phi}(X_i)) \in I(\hat{\phi})]$.

To state a version of Theorem 3, we redefine $\tilde{\Delta}_j$ as at (4.7) with the new definitions of I, $I_1(z)$ and $I_2(x)$. We replace (A5) by the following assumption on the support set $\mathcal{I}$ and the true time transformation ϕ:

(A5′) $\sup_{u\in I_j} \mathrm{mes}[I_j(u, \varphi)\triangle I_j(u, \phi)] \le C \sup_{x\in I_1} |\varphi(x) - \phi(x)|$ for some constant $C > 0$, where $A\triangle B$ denotes the symmetric difference of two sets A and B.

THEOREM 6. *Assume that the conditions of Theorem 4 hold and that conditions* (A2), (A3) *and* (A5′) *are satisfied. Assume also that the bandwidths* g_j *satisfy* $g_j \to 0$ *and* $ng_1g_2/\log n \to \infty$. *If* $\sup_{x\in I_1} |\hat{\phi}(x) - \phi(x)| = O_p(\varepsilon_n)$ *and* $\sup_{(x,y)\in\mathcal{I}} |\hat{p}(x, y) - p(x, y)| = O_p(\varepsilon_n')$, *then* $\hat{f}_j(u) - f_j(u) = \tilde{f}_j(u) - f_j(u) + f_j(u)\tilde{\Delta}_j(u) + O_p(n^{-1/2} + \varepsilon_n^2 + \varepsilon_n'^2 + \varepsilon_n g_1^2 g_2^{-1} + \varepsilon_n g_2 + \varepsilon_n n^{-1/2} g_2^{-3/2}\sqrt{\log n}) + o_p(g_1^2 + g_2^2)$ *uniformly for* $u \in I_j$.

6. Simulation study. For the component functions f_j in the model (2.1), we considered $f_1(u) = 3/2 - u$, $f_2(u) = c(5/4 - 3u^2/4)$. For the function ϕ, we made two choices:

$$\text{Model 1} \qquad \phi(u) = \begin{cases} (u - 1/4)^2 + 15/16, & \text{if } 0 \le u \le 1/2; \\ -(u - 3/4)^2 + 17/16, & \text{if } 1/2 \le u \le 1, \end{cases}$$

$$\text{Model 2} \qquad \phi(u) = 1 - u^2/2.$$

The constant c was chosen so that $\int_{\mathcal{I}} f_1(x) f_2(y\phi(x))\, dx\, dy = 1$, where $\mathcal{I} = \{(x, y) : 0 \le x, y \le 1, x + y \le 1\}$. We generated 500 pseudo sample of sizes $n = 400$ and 1000, from the two models.

For the estimation of ϕ, we computed our estimator on a grid of bandwidth choice $h_1 = h_2$. For each grid point in the bandwidth range, we computed the Monte Carlo estimates of $\mathrm{MISE} = E\int_0^1 (\hat{\phi}(u) - \phi(u))^2\, du$ based on the 500 pseudo samples. We found that, in the first setting, the minimal value of MISE was achieved by the bandwidth choice $h_1 = h_2 = 2.40$ for the sample size $n = 400$, and $h_1 = h_2 = 2.30$ for the sample size $n = 1000$. In the second setting, the bandwidth that gave the minimal MISE was $h_1 = h_2 = 0.90$ for $n = 400$ and $h_1 = h_2 = 0.76$ for $n = 1000$. The panels in Figure 1 depict the boxplots of the values of MISE, ISB and IV computed using the bandwidths on the grids. Here, $\mathrm{ISB} = \int_0^1 (E\hat{\phi}(u) - \phi(u))^2\, du$ and $\mathrm{IV} = \int_0^1 \mathrm{var}(\hat{\phi}(u))\, du$, so that $\mathrm{MISE} = \mathrm{ISB} + \mathrm{IV}$. We report only the results for n = 400 in Figure 1. Those cases with outlying large

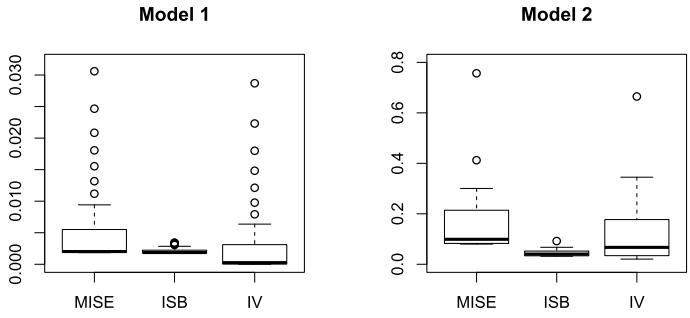

FIG. 1. *Boxplots for the values of MISE, ISB and IV of the estimator* $\hat{\phi}$ *computed using various bandwidth choices, based on* 500 *pseudo samples of size* $n = 400$.

values of MISE for the first model correspond to small bandwidths that produced large values of IV. The results suggest that the variance of the estimator is more influenced by the bandwidth choice than the bias part.

Using the estimates $\hat{\phi}$ based on the bandwidth choices $h_1 = h_2$ that gave the best performance, we computed our estimates of the component functions $\hat{f}_1$ and $\hat{f}_2$. For this estimation, we also took a grid of bandwidth choice $g_1 = g_2$. We computed the mean integrated squared errors

$$\mathrm{MISE}_j = E \int_0^1 \big(\hat{f}_j(u) - f_j(u)\big)^2 \, du, \qquad j = 1, 2$$

with the corresponding values of ISB_j and IV_j. Figure 2 shows the boxplots of the values of MISE_j computed using the bandwidths $g_1 = g_2$ on the grid. Here, we also report the results for $n = 400$ only since the lessons are essentially the same. Comparing the two settings in terms of the accuracy of estimating the component functions f_j, we find that they are not much different. This is because both settings

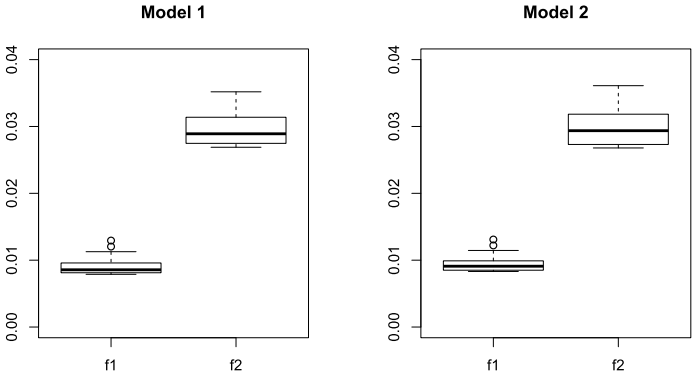

FIG. 2. *Boxplots for the values of* MISE_j *of the estimators* $\hat{f}_j$ *computed using various bandwidth choices of* g *for the two models, based on* 500 *pseudo samples of size* $n = 400$.

have the same component functions and are differ only in the specification of the time transformation ϕ. The results in Figures 1 and 2 suggest that the level of difficulty in the estimation of ϕ does not affect much the accuracy of the estimation of the component functions f_j.

The bandwidth that gave the minimal value of $\mathrm{MISE}_1 + \mathrm{MISE}_2$ in the first setting was $g_1 = g_2 = 0.52$ for $n = 400$ and $g_1 = g_2 = 0.44$ for $n = 1000$. In the case of the second setting, the best performance in terms of $\mathrm{MISE}_1 + \mathrm{MISE}_2$ was achieved by $g_1 = g_2 = 0.50$ for $n = 400$ and $g_1 = g_2 = 0.44$ for $n = 1000$. The values of MISE_j, ISB_j and IV_j for these optimal bandwidths when $n = 400$ are reported in Table 1. Also included in the table are the values of MISE, ISB and IV of $\hat{\phi}$. Although our primary concern is the estimation of the component functions, it is also of interest to see how good the produced two-dimensional density estimator $\hat{f}_1(x)\hat{f}_2(y\hat{\phi}(x))$ behaves. For this, we include in the table the values of MISE, ISB and IV of the two-dimensional estimates computed using the optimal bandwidths. For comparison, we also report the results for the two-dimensional local quadratic estimate defined at (3.1) that does not use the structure of the density. For this local quadratic estimator, we used its optimal bandwidth choice. The results confirm that our two-dimensional density estimator has much better performance than the local quadratic estimator, in both models.

One may be also interested in what happens if one ignores the presence of the nonconstant ϕ and estimates f_j with $\hat{\phi} \equiv 1$, that is, estimates the simple product model $p(x, y) = f_1(x) f_2(y)$, $(x, y) \in \mathcal{I}$. With the corresponding optimal bandwidths, the latter method produced $(\mathrm{MISE}, \mathrm{ISB}, \mathrm{IV}) = (0.0088, 0.0026, 0.0062)$ for f_1 and $(0.0356, 0.0168, 0.0188)$ for f_2 in the case of Model 1, and $(0.0093, 0.0031, 0.0062)$ for f_1 and $(0.0320, 0.0136, 0.0184)$ for f_2 in the case of Model 2. Comparing these with the results in Table 1, we see that estimating ϕ reduced significantly the values of MISE for the second component f_2, which appears to owe to the great reduction in ISB. Note that the accuracy of the estimation

TABLE 1

Mean integrated squared errors (MISE), integrated squared biases (ISB) and integrated variance (IV) of the estimators, based on 500 *pseudo samples of size* $n = 400$

		Component functions			**Joint density** p	
		f_1	f_2	ϕ	**Proposed**	**Local quad.**
Model 1	MISE	0.0080	0.0269	0.0018	0.0137	0.0250
	ISB	0.0025	0.0112	0.0017	0.0037	0.0216
	IV	0.0055	0.0158	0.0001	0.0100	0.0034
Model 2	MISE	0.0085	0.0268	0.0799	0.0144	0.0180
	ISB	0.0027	0.0078	0.0489	0.0039	0.0123
	IV	0.0058	0.0190	0.0310	0.0105	0.0057

TABLE 2
Mean integrated squared errors (MISE), integrated squared biases (ISB) and integrated variance (IV) of the estimators for Model 3 ($\phi \equiv 1$), *based on* 500 *pseudo samples of size* $n = 400$

	Our approach			**Oracle**		
	f_1	f_2	p	f_1	f_2	p
MISE	0.0081	0.0269	0.0136	0.0076	0.0210	0.0136
ISB	0.0025	0.0109	0.0033	0.0022	0.0065	0.0039
IV	0.0056	0.0160	0.0103	0.0054	0.0145	0.0097

of the second component f_2 relies on that of ϕ, and that the method with $\hat{\phi} \equiv 1$ would produce a biased estimator of f_2.

Another thing that is of interest is that how our approach performs when there is no operational time, that is, the true $\phi \equiv 1$. For this, we compared our approach that involves estimating ϕ with the oracle estimators that make use of the knowledge that $\phi \equiv 1$. The results are contained in Table 2. In comparison with the oracle estimators, our approach produced slightly less accurate estimators of the component functions, but gave nearly the same MISE value for the estimator of the joint density function.

We also undertook a sensitivity analysis to check what happens if the structural assumptions of the model are violated, that is, the density of (X, Y) does not consist of three one-dimensional components but is simply a two-dimensional smooth density. For this we generated 500 samples of size $n = 400$ from a bivariate normal distribution with mean $(1/2, 1/2)$ and variance $(1/3, 1/3)$ with correlation $1/2$, but truncated outside the parallelogram $\{(x, y) : -(y/2) + (1/2) \le x \le -(y/2) + 1, 0 \le y \le 1\}$. We compared the local linear and quadratic density estimators of the truncated normal density with the structured estimator that is based on the model (2.1). We found that, with the corresponding optimal bandwidths, the local linear estimator was slightly better than the local quadratic estimator, and it gave (MISE, ISB, IV) $= (0.1015, 0.0861, 0.0154)$, while our structured estimation produced a better result, (MISE, ISB, IV) $= (0.0729, 0.0570, 0.0159)$. This result suggests that the operational time ϕ introduced into the multiplicative density adds a great deal of flexibility to the model so that it approximates quite well densities violating the independence assumption.

In practical implementation of our method, one may employ a K-fold cross-validation criterion to choose the bandwidths h and g. To be specific, one splits the whole dataset into K (nearly) equal parts, $\{(X_i, Y_i) : i \in J_k\}$, $1 \le k \le K$. For each partition J_k, one computes

$$\mathrm{CV}_k(h, g) = \int_{\mathcal{I}} \hat{p}_{h,g,-k}(x, y)^2 \, dx \, dy - \frac{2}{|J_k|} \sum_{i \in J_k} \hat{p}_{h,g,-k}(X_i, Y_i),$$

TABLE 3
Mean integrated squared errors (MISE), integrated squared biases (ISB) and integrated variance (IV) of the estimators with tenfold cross-validated bandwidths, based on 500 pseudo samples of size $n = 400$

	Model 1		**Model 2**		**Model 3**	
	f_1	f_2	f_1	f_2	f_1	f_2
MISE	0.0074	0.0262	0.0078	0.0257	0.0075	0.0269
ISB	0.0028	0.0126	0.0033	0.0101	0.0030	0.0133
IV	0.0046	0.0136	0.0045	0.0156	0.0045	0.0136

where $|J_k|$ is the size of the index set J_k and $\hat{p}_{h,g,-k}$ denotes our structured density estimate computed from the dataset with the kth partition being deleted, that is, from $\{(X_i, Y_i) : i \notin J_k\}$, based on the bandwidth choice (h, g). The above CV criterion is common in density estimation; see Park and Marron (1990), for example. It is an estimate of $\int_{\mathcal{I}}(\hat{p}_{h,g,-k}(x, y) - p(x, y))^2\,dx\,dy +$ (irrelevant term). The K-fold cross-validated choice is then defined by

$$(h_{\mathrm{cv}}, g_{\mathrm{cv}}) = (\hat{h}, \hat{g}) \times (1 - 1/K)^{1/5},$$

where $(\hat{h}, \hat{g})$ is the minimizer of $\mathrm{CV}(h, g) = \sum_{k=1}^{K} \mathrm{CV}_k(h, g)/K$. Note that the correction factor $(1 - 1/K)^{1/5}$ is needed since $(\hat{h}, \hat{g})$ is suitable for the sample size $n(1 - 1/K)$ rather than n.

To see how K-fold cross-validated bandwidths perform in this particular problem, we chose $K = 10$ and applied the method to the three models. The results are summarized in Table 3. Comparing the results with those in Tables 1 and 2, we find that the cross-validated bandwidth selector works fairly well, giving comparable performance with the MISE-optimal bandwidth. Motivated by this good performance, we used the tenfold cross-validated bandwidth in our data example in Section 7.

7. Motor insurance data. As an example of implementing our method, we considered reported and outstanding claims from a motor insurance business line in Cyprus. For each claim, the dataset includes (EntryDate), (ClaimStatus) and (StatusDate). (EntryDate) is the date the claim was reported and entered the system, (StatusDate) is the date of the last update of (ClaimStatus) that has three categories: P for paid and settled; W for not paid but settled; O for open and not settled. Among 58,453 claims reported during the period January 12, 2004, to July 31, 2014, those claims with status O were deleted since for these claims the date of settlement was not observed. The number of deleted claims was 1865, and thus the number of the claims that we used to fit our model was 56,588.

In this example, (EntryDate) corresponds to the variable X, and the delay time until settlement, (StatusDate)−(EntryDate), to the variable Y. To apply our

model (2.1) with $\mathcal{I} = \{(x, y) : x \geq 0, y \geq 0, x + y \leq 1\}$, we transformed the daily claim data in the following way. We first enumerated the calendar dates from 1 to 3854, with 1 corresponding to January 12, 2004, and 3854 to July 31, 2014, and then changed (EntryDate) and (StatusDate) to the respective integers on the new discrete scale. This would result in a dataset for the variables [(EntryDate), (StatusDate) − (EntryDate)] on the discrete triangular $\{(j, k) : 1 \leq j \leq 3854, 0 \leq k \leq 3853\}$. We then transformed them to (X, Y) by

$$X = \frac{\text{(EntryDate)} - 1 + U_1}{3854}, \qquad Y = \frac{\text{(StatusDate)} - \text{(EntryDate)} + U_2}{3854},$$

where (U_1, U_2) is a two-dimensional uniform random variate on the unit square $[0, 1]^2$. Here, the perturbation by uniform random variates is done to make the converted data (X, Y) take values on the two-dimensional continuous time scale. This gives a converted dataset $\{(X_i, Y_i) : 1 \leq i \leq 56{,}588\}$. We applied to this dataset our method of estimating the structured density p of (X, Y).

We took a common bandwidth $h = h_1 = h_2$ for the estimation of the time transformation ϕ, and a common bandwidth $g = g_1 = g_2$ for the estimation of the component functions. We selected (h, g) by the tenfold cross-validated criterion described in Section 6 ($K = 10$).

The results of the application of our method to the insurance claim data are shown in Figure 3. In the left panel, the solid curve depicts the estimate of the time transformation ϕ and the dashed (dotted) is a 90% (95%) pointwise bootstrap confidence band for ϕ. The $100(1 - \alpha)$ confidence bands $[2\hat{\phi}(x) - U_\alpha(x), 2\hat{\phi}(x) - L_\alpha(x)]$ were based on 1000 bootstrap samples, where $L_\alpha(x)$ and $U_\alpha(x)$ are the bootstrap estimates of the $\alpha/2$ and $(1 - \alpha/2)$ quantiles, respectively, of the distribution of $\hat{\phi}(x)$. We note that the confidence bands are narrowed down to the

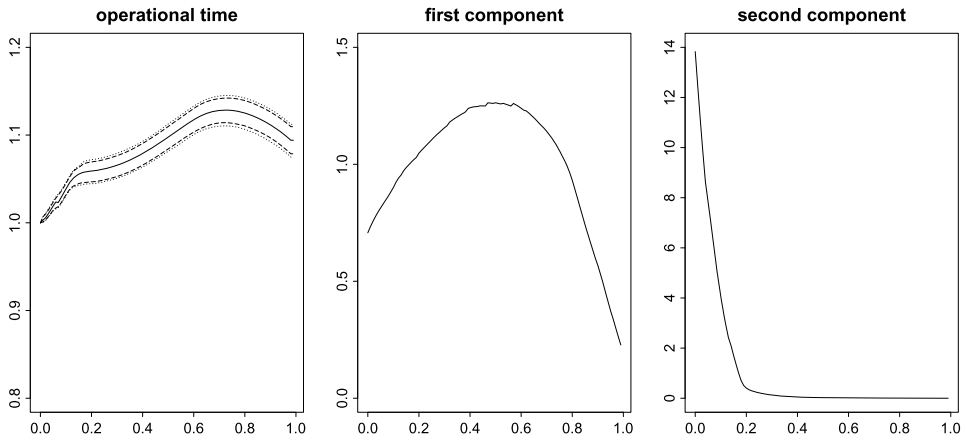

FIG. 3. *The estimate of the time transformation ϕ with* 90% (*dashed*) *and* 95% (*dotted*) *pointwise confidence bands* (*left*), *the estimates of the first component function f_1* (*middle*) *and the second component function f_2* (*right*), *obtained by applying the model* (2.1) *to the insurance claim data.*

point $\hat{\phi}(x) = 1$ at $x = 0$ because of our normalization $\hat{\phi}(0) = 1 = \phi(0)$, see (3.2). The bootstrap confidence bands indicate that the underlying transformation ϕ is not constant, so that the model (2.1) does not degenerate to the simple product model $p(x, y) = f_1(x) f_2(y)$ considered by Mammen, Martínez Miranda and Nielsen (2015). The estimated ϕ suggests that the speed of time has an increasing tendency and that speed of time has increased by around 10% over the 10-year period considered. This is more or less in line with intuitive expectations to the model on how much improved technology has speeded up the process of getting incidents of claims settled. The decline of $\hat{\phi}$ after its peak might be because the company overall has decreased the number of employees when it saw the benefits of the advanced technology. The estimate of the first component measures business exposure, thus the middle panel of Figure 3 indicates that the business line had increasing exposure in the first half of the period, but ran out later and perhaps was replaced by new products recorded in a separate dataset. The second component that measures time to settlement follows more or less the usual pattern known from motor insurance business lines that the claims development is quite fast.

One may use our estimated model to forecast the density on an unobserved area. In general, let S be a subset of $[0, 1]^2$, outside of the observed area $\mathcal{I}$, where one wants to forecast the density. With the estimated density model $\hat{p}(x, y) = \hat{f}_1(x) \hat{f}_2(y\hat{\phi}(x))$, the relative mass of the probability on S with respect to that on $\mathcal{I}$ is estimated by

$$A(S) = \int_S \hat{f}_1(x) \hat{f}_2\big(y\hat{\phi}(x)\big)\, dx\, dy. \tag{7.1}$$

The number of future observations that fall in the area S is then forecasted by $N(S) = n \cdot A(S)$, where n is the sample size, that is, the total number of observations in $\mathcal{I}$. To apply the forecasting method to the motor insurance dataset and evaluate its accuracy, we re-estimated the model (2.1) now using the data observed until the year 2012. We forecasted the number of claims settled in the year 2013 according to the formula at (7.1). The actual number was 4547. Our approach produced 4487, while the forecasting based on the simple product model $p(x, y) = f_1(x) f_2(y)$ gave 4226.

APPENDIX

A.1. Proof of Theorem 2. In the following proof, we will use the symbol W to denote functions that have bounded continuous partial derivatives, and W^* for continuous bounded functions. The symbols will be used for different functions, even in the same formula. They will denote univariate functions and bivariate functions as well. Furthermore, for simplicity of notation, we assume that $h_1 = h_2 = h$.

Put $\Delta(x) = \log \hat{\phi}(x) - \log \hat{\phi}(h) - [\log \phi(x) - \log \phi(h)]$. We will show that

$$\Delta(x) = O_p\big(n^{-2/5}\big), \qquad 0 \le x < 1. \tag{A.1}$$

It can be shown by slightly modified and simpler arguments that

$$\log \hat{\phi}(h) - \log \phi(h) = O_p\big(n^{-2/5}\big). \tag{A.2}$$

The bounds (A.1) and (A.2) imply (3.5). For a proof of (3.6), one may show, instead of (A.1) and (A.2), the slightly stronger claim

$$\sup_{x \in [0,1-\epsilon]} E\Delta^2(x) = O\big(n^{-4/5}\big). \tag{A.3}$$

This can be done by a slightly more careful use of the arguments in the proof of (A.1). For a proof of (3.7), one makes use of exponential inequalities for the terms of the stochastic expansion that we will consider below in the proof of (A.1).

We now come to the proof of (A.1). By Taylor's expansion, one gets that

$$\begin{aligned} \Delta(x) &= \int_h^x \bigg[\frac{\hat{G}_{11}(u) - \hat{G}_{10}(u)\hat{G}_{01}(u)}{\hat{G}_{02}(u) - \hat{G}_{01}(u)^2} \\ &\qquad - \frac{G_{11}(u) - G_{10}(u)G_{01}(u)}{G_{02}(u) - G_{01}(u)^2} \bigg] du \\ &= \Delta_1(x) + \Delta_2(x) + \Delta_3(x) + R(x), \end{aligned} \tag{A.4}$$

where Δ_1 comprises all linear terms of the form $\int_h^x W(u)(\hat{G}_{jk}(u) - G_{jk}(u))\,du$ of a Taylor expansion of the integrand of the integral in (A.4). The second term Δ_2 collects those terms of quadratic order, $\int_h^x W(u)(\hat{G}_{jk}(u) - G_{jk}(u))(\hat{G}_{j'k'}(u) - G_{j'k'}(u))\,du$, and Δ_3 contains all cubic terms. Among these linear, quadratic and cubic terms, the most complex terms are those that involve $\hat{G}_{11}$. Note that $\hat{G}_{11}$ contains a product of two partial derivatives, whereas $\hat{G}_{jk}$ for $(j,k) \neq (1,1)$ includes at most one partial derivative. For the remainder term R, it holds that $R(x) = O_p(n^{-2/5})$. This bound follows from

$$E\big(\hat{G}_{jk}(u) - G_{jk}(u)\big)^4 = O\big(n^{-2/5}\big)$$

and a bound on the variance of $\int_h^x (\hat{G}_{jk}(u) - G_{jk}(u))^4\,du$. One can show that the bound on R holds uniformly for $0 \le x \le 1 - \epsilon$.

We now prove

$$\int_h^x W(u)\big(\hat{G}_{11}(u) - G_{11}(u)\big)\,du = O_p\big(n^{-2/5}\big). \tag{A.5}$$

Using the same arguments as for the proof of (A.5), one can show that the other terms of Δ_1 are of order $O_p(n^{-2/5})$. This implies that

$$\Delta_1(x) = O_p\big(n^{-2/5}\big). \tag{A.6}$$

For the proof of (A.5), we redefine the vector $\mathbf{a}(u, v; x, y)$ in Section 3 as

$$\begin{aligned} \mathbf{a}\big(u', v'; u, v\big) = \big(&1, (u'-u)^2/h^2, (v'-v)^2/h^2, \\ &(u'-u)/h, (v'-v)/h, (u'-u)(v'-v)/h^2\big)^\top \end{aligned}$$

and also redefine $\hat{\mathbf{b}}(u,v)$ and $\mathbf{A}(u,v)$ in accordance with this change. In this way, it is easier to see how the inverse matrix $\mathbf{A}^{-1}(u,v)$ looks like. Indeed, for (u,v) in the interior region $\mathcal{I}_0$,

$$\mathbf{A}(u,v)=\begin{pmatrix}1&\nu_2&\nu_2&0&0&0\\ \nu_2&\nu_4&\nu_2^2&0&0&0\\ \nu_2&\nu_2^2&\nu_4&0&0&0\\ 0&0&0&\nu_2&0&0\\ 0&0&0&0&\nu_2&0\\ 0&0&0&0&0&\nu_2^2\end{pmatrix},$$

where $\nu_j=\int_{-1}^{1}z^jK(z)\,dz$ are the complete moments of K. Note that the interior of $\mathcal{I}$ in our problem is given by

$$\mathcal{I}_0=\{(x,y):x\geq h_1,\,y\geq h_2,\,x+y\leq 1-h_1-h_2\}.$$

From the structure of $\mathbf{A}(u,v)$ for $(u,v)\in\mathcal{I}_0$ we get, for example,

$$\begin{aligned}\hat{\eta}_{10}(u,v)&=\big(\text{the fourth entry of }\mathbf{A}^{-1}(u,v)\hat{\mathbf{b}}(u,v)\big)\\ &=\nu_2^{-1}n^{-1}h^{-2}\sum_{i=1}^{n}\left(\frac{X_i-u}{h}\right)K\left(\frac{X_i-u}{h}\right)K\left(\frac{Y_i-v}{h}\right).\end{aligned}\tag{A.7}$$

Also, from the standard kernel smoothing theory we obtain that, now uniformly for $(x,y)\in\mathcal{I}$,

$$E\big(\mathbf{A}^{-1}(x,y)\hat{\mathbf{b}}(x,y)\big)-\boldsymbol{\eta}(x,y)=o\big(h^2\big),\tag{A.8}$$

where $\boldsymbol{\eta}=(p,h^2p_{20},h^2p_{02},hp_{10},hp_{01},h^2p_{11})^\top$ and $p_{jk}(x,y)=\partial^{j+k}p(x,y)/\partial x^j\,\partial y^k$. The bound (A.8) follows directly from

$$\begin{aligned}&\int_{\mathcal{I}}\mathbf{a}(u,v;x,y)h^{-2}K\left(\frac{u-x}{h}\right)K\left(\frac{v-y}{h}\right)\big(p(u,v)-\mathbf{a}(u,v;x,y)^\top\boldsymbol{\eta}(x,y)\big)\,du\,dv\\ &\quad=o\big(h^2\big).\end{aligned}$$

Now, for the proof of (A.5) note that

$$\begin{aligned}&\int_h^x W(u)\big(\hat{G}_{11}(u)-G_{11}(u)\big)\,du\\ &\quad=\int_h^x\int_0^{1-u}W(u,v)\big(h^{-1}\hat{\eta}_{10}(u,v)-p_{10}(u,v)\big)\,dv\,du\\ &\qquad+\int_h^x\int_0^{1-u}W(u,v)\big(h^{-1}\hat{\eta}_{01}(u,v)-p_{01}(u,v)\big)\,dv\,du\\ &\qquad+\int_h^x\int_0^{1-u}W(u,v)\big(h^{-1}\hat{\eta}_{10}(u,v)-p_{10}(u,v)\big)\\ &\qquad\times\big(h^{-1}\hat{\eta}_{01}(u,v)-p_{01}(u,v)\big)\,dv\,du\\ &\qquad+R^*(x)+O_p\big(n^{-2/5}\big),\end{aligned}\tag{A.9}$$

where R^* comprises integrals of products containing the factor $(\hat{\eta}_{00}(u,v) - p(u,v))$. These terms can be analysed by standard kernel smoothing techniques and they are all of order $O_p(n^{-2/5})$.

We prove that the first three terms on the right-hand side of (A.9) are of order $O_p(n^{-2/5})$. For the study of the first term, we claim that

$$\int_h^x \int_0^h W(u,v)\big(h^{-1}\hat{\eta}_{10}(u,v) - p_{10}(u,v)\big)\,dv\,du = O_p\big(n^{-2/5}\big), \tag{A.10}$$

$$\int_h^x \int_{1-u-2h}^{1-u} W(u,v)\big(h^{-1}\hat{\eta}_{10}(u,v) - p_{10}(u,v)\big)\,dv\,du = O_p\big(n^{-2/5}\big). \tag{A.11}$$

For the proof of the claim (A.10), note that $\int_h^x W(u,v)(h^{-1}\hat{\eta}_{10}(u,v) - p_{10}(u,v))\,du$ behaves like the error of a one-dimensional kernel derivative estimator and it is thus of order $O_p(n^{-1/2}h^{-3/2} + h)$. Integration of the latter over the interval $[0,h]$ gives (A.10). The claim (A.11) can be verified similarly. Thus, for getting that the first term at (A.9) is of order $O_p(n^{-2/5})$ it remains to show that

$$\begin{aligned}\int_h^x \int_h^{1-u-2h} & W(u,v)\big(h^{-1}\hat{\eta}_{10}(u,v) - p_{10}(u,v)\big)\,dv\,du \\ &= O_p\big(n^{-2/5}\big).\end{aligned} \tag{A.12}$$

For the proof of the claim (A.12), we make use of the expression (A.7) for the estimator of $hp_{01}(u,v)$. We observe $h^{-1}\hat{\eta}_{10}(u,v) = \partial\tilde{p}(u,v)/\partial u$ with

$$\begin{aligned}\tilde{p}(u,v) &= n^{-1}\sum_{i=1}^n h^{-2}L\left(\frac{X_i - u}{h}\right)K\left(\frac{Y_i - v}{h}\right), \\ L(v) &= -\nu_2^{-1}\int_{-1}^v zK(z)\,dz.\end{aligned} \tag{A.13}$$

Using that W has bounded continuous partial derivatives, we get by changing the order of integration and by integration-by-part that

$$\begin{aligned}&\int_h^x \int_h^{1-u-2h} W(u,v)\big(h^{-1}\hat{\eta}_{10}(u,v) - p_{10}(u,v)\big)\,dv\,du \\ &\quad= \int_h^{1-3h}\int_h^{x\wedge(1-v-2h)} W(u,v)\big(\partial\tilde{p}(u,v)/\partial u - p_{10}(u,v)\big)\,du\,dv \\ &\quad= \int_h^{1-3h} W(u,v)\big(\tilde{p}(u,v) - p(u,v)\big)\Big|_{u=h}^{u=x\wedge(1-v-2h)}\,dv \\ &\qquad- \int_h^{1-3h}\int_h^{x\wedge(1-v-2h)} W^*(u,v)\big(\tilde{p}(u,v) - p(u,v)\big)\,du\,dv.\end{aligned} \tag{A.14}$$

For the first term on the right-hand side of the second equation of (A.14) we get that it behaves like a one-dimensional kernel estimator because the two-dimensional

kernel estimator $\tilde{p}(u,v)$, defined at (A.13), is integrated out along a line. Thus, the first term is of order $O_p(n^{-2/5})$. Because the second term is also of order $O_p(n^{-2/5})$, we establish (A.12).

From the arguments in the preceding two paragraphs, we conclude that the first term on the right-hand side of (A.9) is of order $O_p(n^{-2/5})$. By similar arguments, one can show that the second term is also of order $O_p(n^{-2/5})$. For the treatment of the third term, we will show that

$$\int_h^x \int_h^{1-u-2h} W(u,v)\big(h^{-1}\hat{\eta}_{10}(u,v) - p_{10}(u,v)\big) \times \big(h^{-1}\hat{\eta}_{01}(u,v) - p_{01}(u,v)\big)\,dv\,du = O_p\big(n^{-2/5}\big). \tag{A.15}$$

By the consideration of additional boundary terms of the third terms, we conclude that the third term is also of order $O_p(n^{-2/5})$. Thus, for (A.5) it remains to prove (A.15). This also completes the proof of (A.6).

For the proof of (A.15), we put

$$\bar{p}(u,v) = n^{-1}\sum_{i=1}^{n} h^{-2} K\left(\frac{X_i - u}{h}\right) L\left(\frac{Y_i - v}{h}\right).$$

Note that $h^{-1}\hat{\eta}_{01}(u,v) = \partial\bar{p}(u,v)/\partial v$. Thus, the left-hand side of (A.15) equals $\sum_{k=1}^{4} J_k(x)$, where

$$\begin{aligned}
J_1(x) &= \int_h^x \int_h^{1-u-2h} W(u,v)\left(\frac{\partial\tilde{p}(u,v)}{\partial u} - E\frac{\partial\tilde{p}(u,v)}{\partial u}\right) \\
&\quad\times \left(\frac{\partial\bar{p}(u,v)}{\partial v} - E\frac{\partial\bar{p}(u,v)}{\partial v}\right) dv\,du, \\
J_2(x) &= \int_h^x \int_h^{1-u-2h} W(u,v)\left(\frac{\partial\tilde{p}(u,v)}{\partial u} - E\frac{\partial\tilde{p}(u,v)}{\partial u}\right) \\
&\quad\times \left(E\frac{\partial\bar{p}(u,v)}{\partial v} - p_{01}(u,v)\right) dv\,du, \\
J_3(x) &= \int_h^x \int_h^{1-u-2h} W(u,v)\left(E\frac{\partial\tilde{p}(u,v)}{\partial u} - p_{10}(u,v)\right) \\
&\quad\times \left(\frac{\partial\bar{p}(u,v)}{\partial v} - E\frac{\partial\bar{p}(u,v)}{\partial v}\right) dv\,du, \\
J_4(x) &= \int_h^x \int_h^{1-u-2h} W(u,v)\left(E\frac{\partial\tilde{p}(u,v)}{\partial u} - p_{10}(u,v)\right) \\
&\quad\times \left(E\frac{\partial\bar{p}(u,v)}{\partial v} - p_{01}(u,v)\right) dv\,du.
\end{aligned}$$

It holds that $J_4(x) = O(n^{-2/5})$ because of the fact $E\partial\tilde{p}(u,v)/\partial u - p_{10}(u,v) = O(n^{-1/5})$ and $E\partial\bar{p}(u,v)/\partial v - p_{01}(u,v) = O(n^{-1/5})$, uniformly for $(u,v) \in \mathcal{I}$.

For J_2, we get that

$$\begin{aligned} J_2(x) = & -\int_h^x \int_h^{1-u-2h} W^*(u,v) n^{-1} h^{-2} \sum_{i=1}^n \Big[L'\Big(\frac{X_i-u}{h}\Big) K\Big(\frac{Y_i-v}{h}\Big) \\ & - EL'\Big(\frac{X_i-u}{h}\Big) K\Big(\frac{Y_i-v}{h}\Big)\Big] dv\, du \\ = & \, n^{-1} \sum_{i=1}^n \big(W_n(X_i,Y_i;x) - EW_n(X_i,Y_i;x)\big) \end{aligned}$$

for some bounded function W_n. Thus, we have $J_2(x) = O_p(n^{-1/2})$. The same holds for J_3. Therefore, for (A.15) it remains to show $J_1(x) = O_p(n^{-2/5})$. For the proof of this claim, we let

$$\begin{aligned} R_{n,ij}(x) = & \, h^{-2} \int_h^x \int_h^{1-u-2h} W(u,v) \Big[L'\Big(\frac{X_i-u}{h}\Big) K\Big(\frac{Y_i-v}{h}\Big) \\ & - EL'\Big(\frac{X_i-u}{h}\Big) K\Big(\frac{Y_i-v}{h}\Big)\Big] \cdot \Big[K\Big(\frac{X_j-u}{h}\Big) L'\Big(\frac{Y_j-v}{h}\Big) \\ & - EK\Big(\frac{X_j-u}{h}\Big) L'\Big(\frac{Y_j-v}{h}\Big)\Big] dv\, du. \end{aligned}$$

Then we can write

$$J_1(x) = n^{-2}h^{-4} \sum_{1\le i\ne j\le n} R_{n,ij} + n^{-2}h^{-4} \sum_{i=1}^n R_{n,ii} = J_{1a} + J_{1b}.$$

Also, put

$$\begin{aligned} R^*_{n,ij}(x) = & \, h^{-2} \int_h^x \int_h^{1-u-2h} W(u,v) K\Big(\frac{X_i-u}{h}\Big) K\Big(\frac{Y_j-v}{h}\Big) \\ & \times L'\Big(\frac{X_j-u}{h}\Big) L'\Big(\frac{Y_i-v}{h}\Big) dv\, du. \end{aligned}$$

For $i \ne j$, the random variable $R^*_{n,ij}$ is bounded and satisfies

$$R^*_{n,ij}(x) = 0 \qquad \text{if } |X_i - X_j| \ge 2h \text{ or } |Y_i - Y_j| \ge 2h.$$

By the definition of J_{1a}, we get by using a simple inequality for second moments of U-statistics that

$$EJ_{1a}^2 \le n^{-4}h^{-8} \cdot 2 \cdot \sum_{1\le i\ne j\le n} ER^{*2}_{n,ij} = O\big(n^{-2}h^{-8}h^2\big) = O\big(n^{-4/5}\big).$$

This gives $J_{1a} = O_p(n^{-2/5})$. It remains to check $J_{1b} = O_p(n^{-2/5})$. For checking this claim, we note that $h^{-1}R_{n,ii}(x)I((X_i,Y_i) \in \mathcal{I}_*(x))$ is a bounded random variable, where $\mathcal{I}_*(x) = \{(u,v) \in \mathcal{I} : 2h \le u \le x-h, 2h \le v, u+v \le 1-4h\}$. This

follows from the fact $\int_{-1}^{1} K(u)L'(u)\,du = 0$ and from an expansion of $W(u, v)$ around $u = X_i$ and $v = Y_i$. Furthermore, we have that $R_{n,ii}$ is a bounded random variable and that $P[(X_i, Y_i) \in \mathcal{I}_{**}(x) - \mathcal{I}_*(x)] = O(h)$, where $\mathcal{I}_{**}(x) = \{(u', v') \in \mathcal{I} : (u' - u, v' - v) \in [-h, h]^2$ for some (u, v) with $h \le u \le x, h \le v \le 1 - u - 2h\}$. These properties of $R_{n,ii}$ can be used to show $J_{1b}(x) = O_p(n^{-2/5})$. This completes the proof of (A.5).

For the statement of the theorem, it remains to check that $\Delta_2(x) = O_p(n^{-2/5})$ and $\Delta_3(x) = O_p(n^{-2/5})$. The study of Δ_2 leads to quadratic terms that are similar to (A.15). All these terms can be treated as in the study of Δ_1. Additionally, we will have terms of the type

$$\int_h^x \int_h^{1-u-2h} \int_h^{1-u-2h} W(u, v)\big(h^{-1}\hat{\eta}_{01}(u, v) - p_{01}(u, v)\big) \times W\big(u, v'\big)\big(h^{-1}\hat{\eta}_{10}\big(u, v'\big) - p_{10}\big(u, v'\big)\big)\,dv\,dv'\,du.$$

Because of the additional integration, the analysis of these terms is much easier than the study of (A.15). The same argument applies to other terms of Δ_2 and Δ_3. By lengthy but simple calculations, one may get $\Delta_2(x) = O_p(n^{-2/5})$ and $\Delta_3(x) = O_p(n^{-2/5})$.

A.2. Proof of Theorem 5. Theorem 5 may be proved along the lines of the proof of Theorem 5. To list the essential changes, the interior $\mathcal{I}_0$ is now given in a general form by

$$\mathcal{I}_0 = \left\{(x, y) \in \mathcal{I} : \left\{\left(\frac{u - x}{h_1}, \frac{v - y}{h_2}\right) : (u, v) \in \mathcal{I}\right\} \supset [-1, 1]^2\right\}$$

and $\hat{\phi}(h)$ and $\phi(h)$ in the definition of $\Delta(x)$ at (A.1) are replaced by $\hat{\phi}(x_{\min,h})$ and $\phi(x_{\min,h})$, respectively, where $x_{\min,h} = \min\{x : (x, y) \in \mathcal{I}_0$ for some $y\}$. The integrals over the interval $[h, x]$ at (A.4) and (A.5) are now over $[x_{\min,h}, x]$, the integration at (A.9) needs to be over the set $\{(u, v) : x_{\min,h} \le u \le x, v \in \mathcal{I}_2(u)\}$, the two integrals at (A.10) and (A.11) are put together to be the integral over $\{(u, v) : x_{\min,h} \le u \le x, (u, v) \in \mathcal{I}_0^c\}$, and the integrals at (A.12), (A.14) and (A.15) should be over $\mathcal{I}_0(x) \equiv \{(u, v) : x_{\min,h} \le u \le x, (u, v) \in \mathcal{I}_0\}$. The sets $\mathcal{I}_*$ and $\mathcal{I}_{**}$ at the end of the proof are redefined as $\mathcal{I}_*(x) = \{(u, v) \in \mathcal{I} : [u - h, u + h] \times [v - h, v + h] \subset \mathcal{I}_0(x)\}$ and $\mathcal{I}_{**}(x) = \{(u', v') \in \mathcal{I} : (u' - u, v' - v) \in [-h, h]^2$ for some $(u, v) \in \mathcal{I}_0(x)\}$.

SUPPLEMENTARY MATERIAL

Supplement to "Operational time and in-sample density forecasting" (DOI: 10.1214/16-AOS1486SUPP; .pdf). We provide the proofs of Theorems 3 and 6 in the supplement.

REFERENCES

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York. MR1198884

BARAUD, Y. and BIRGÉ, L. (2014). Estimating composite functions by model selection. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 285–314. MR3161532

CHENG, M.-Y. (1997). A bandwidth selector for local linear density estimators. *Ann. Statist.* **25** 1001–1013. MR1447738

FAN, J., HECKMAN, N. E. and WAND, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90** 141–150. MR1325121

HOROWITZ, J. L. and MAMMEN, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Ann. Statist.* **35** 2589–2619. MR2382659

JIANG, J., FAN, Y. and FAN, J. (2010). Estimation in additive models with highly or nonhighly correlated covariates. *Ann. Statist.* **38** 1403–1432. MR2662347

JUDITSKY, A. B., LEPSKI, O. V. and TSYBAKOV, A. B. (2009). Nonparametric estimation of composite functions. *Ann. Statist.* **37** 1360–1404. MR2509077

KUANG, D., NIELSEN, B. and NIELSEN, J. P. (2008a). Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika* **95** 979–986. MR2461224

KUANG, D., NIELSEN, B. and NIELSEN, J. P. (2008b). Forecasting with the age-period-cohort model and the extended chain-ladder model. *Biometrika* **95** 987–991. MR2461225

KUANG, D., NIELSEN, B. and NIELSEN, J. P. (2009). Chain-ladder as maximum likelihood revisited. *Annals of Actuarial Science* **4** 105–121.

KUANG, D., NIELSEN, B. and NIELSEN, J. P. (2011). Forecasting in an extended chain-ladder-type model. *J. Risk Insur.* **78** 345–359.

LEE, R. D. and CARTER, L. R. (1992). Modeling and forecasting U.S. mortality. *J. Amer. Statist. Assoc.* **87** 659–671.

LEE, Y. K., MAMMEN, E. and PARK, B. U. (2010). Backfitting and smooth backfitting for additive quantile models. *Ann. Statist.* **38** 2857–2883. MR2722458

LEE, Y. K., MAMMEN, E. and PARK, B. U. (2012). Flexible generalized varying coefficient regression models. *Ann. Statist.* **40** 1906–1933. MR3015048

LEE, Y. K., MAMMEN, E., NIELSEN, J. P. and PARK, B. U. (2015). Asymptotics for in-sample density forecasting. *Ann. Statist.* **43** 620–651. MR3319138

LEE, Y. K., MAMMEN, E., NIELSEN, J. P. and PARK, B. U. (2016). Supplement to "Operational time and in-sample density forecasting." DOI:10.1214/16-AOS1486SUPP.

MAMMEN, E., MARTÍNEZ MIRANDA, M. D. and NIELSEN, J. P. (2015). In-sample forecasting applied to reserving and mesothelioma mortality. *Insurance Math. Econom.* **61** 76–86. MR3324046

MAMMEN, E. and NIELSEN, J. P. (2003). Generalised structured models. *Biometrika* **90** 551–566. MR2006834

MAMMEN, E., PARK, B. U. and SCHIENLE, M. (2014). Additive models: Extensions and related models. In *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics* (J. S. Racine, L. Su and A. Ullah, eds.) 176–211. Oxford Univ. Press, Oxford. MR3306926

MARTÍNEZ-MIRANDA, M. D., NIELSEN, J. P., SPERLICH, S. and VERRALL, R. J. (2013). Continuous chain ladder: Reformulating and generalising a classical insurance problem. *Expert Syst. Appl.* **40** 5588–5603.

MARTÍNEZ MIRANDA, M. D., NIELSEN, J. P. and VERRALL, R. (2012). Double chain ladder. *Astin Bull.* **42** 59–76. MR2931855

MIKOSCH, T. (2009). *Non-life Insurance Mathematics*, 2nd ed. *Universitext*. Springer, Berlin. MR2503328

PARK, B. U. and MARRON, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85** 66–72.

REID, P. H. (1978). Claims reserves in general insurance. *J. Inst. Actuar.* **105** 211–296.

RUPPERT, D. and WAND, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22** 1346–1370. MR1311979

TAYLOR, G. C. (1981). Speed finalisation of claims and claims run-off analysis. *Astin Bull.* **12** 81–100.

TAYLOR, G. C. (1982). Zehnwirth's comment on the see-saw method: A reply. *Insurance Math. Econom.* **1** 105–108.

WILKE, R. (2016). Forecasting macroeconomic labour market flows: What can we learn from micro level analysis? (submitted manuscript).

YU, K., PARK, B. U. and MAMMEN, E. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.* **36** 228–260. MR2387970

ZEHNWIRTH, B. (1982). Comments on Taylor's see-saw approach to claims reserving. *Insurance Math. Econom.* **1** 99–103.

ZHANG, X., PARK, B. U. and WANG, J.-L. (2013). Time-varying additive models for longitudinal data. *J. Amer. Statist. Assoc.* **108** 983–998. MR3174678

Y. K. LEE
DEPARTMENT OF STATISTICS
KANGWON NATIONAL UNIVERSITY
CHUNCHEON 200-701
KOREA
E-MAIL: youngklee@kangwon.ac.kr

J. P. NIELSEN
CASS BUSINESS SCHOOL
CITY UNIVERSITY LONDON
106 BUNHILL ROW
LONDON EC1Y 8TZ
UNITED KINGDOM
E-MAIL: Jens.Nielsen.1@city.ac.uk

E. MAMMEN
INSTITUTE FUER ANGEWANDTE MATHEMATIK
UNIVERSITAET HEIDELBERG
IM NEUENHEIMER FELD 294
69120 HEIDELBERG
GERMANY
E-MAIL: mammen@math.uni-heidelberg.de

B. U. PARK
DEPARTMENT OF STATISTICS
SEOUL NATIONAL UNIVERSITY
SEOUL 151-747
KOREA
E-MAIL: bupark@stats.snu.ac.kr